Supplementary Information

1 Bispectrum

An arbitrary function ρ defined on the surface of a 4D sphere can be numerically represented using the hyperspherical harmonic functions $U^j_{m'm}(\phi,\theta,\theta_0)$. The hyperspherical harmonics form an orthonormal basis set thus ρ can be represented as

$$\rho = \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} c^j_{m'm} U^j_{m'm}.$$

The expansion coefficients $c^j_{m'm}$ can be calculated via

$$c^j_{m'm} = \langle U^j_{m'm} | \rho \rangle,$$

where $\langle . | . \rangle$ denotes the inner product. For clarity, the vectors $\mathbf{c}^j$ are constructed from the expansion coefficients $c^j_{m'm}$. A unitary operation $\hat{R}$, such as a rotation, acting on ρ transforms the coefficient vectors $\mathbf{c}^j$ according to

$$\mathbf{c}'^j = \mathbf{R}^j \mathbf{c}^j,$$

where $\mathbf{R}^j$ are unitary matrices, i.e. $\left(\mathbf{R}^j\right)^\dagger \mathbf{R}^j = \mathbf{I}$.

The direct product of two rotational matrices, $\mathbf{R}^{j_1}$ and $\mathbf{R}^{j_2}$ can be decomposed into a direct product of $\mathbf{R}^j$ matrices by a unitary transformation

$$\mathbf{R}^{j_1} \otimes \mathbf{R}^{j_2} = \left(\mathbf{H}^{j_1,j_2}\right)^\dagger \left[\bigoplus_{j=|j_1-j_2|}^{j_1+j_2} \mathbf{R}^j \right] \mathbf{H}^{j_1,j_2},$$

where the matrix $\mathbf{H}^{j_1 j_2}$ is the four-dimensional analouge of the Clebsch-Gordan coefficients. In fact, the elements of the matrix are obtained as product of Clebsch-Gordan coefficients: $H^{lmm'}_{l_1 m_1 m'_1, l_2 m_2 m'_2} \equiv C^{lm}_{l_1 m_1 l_2 m_2} C^{lm'}_{l_1 m'_1 l_2 m'_2}$. The direct product of the coefficient vectors $\mathbf{c}^{j_1}$ and $\mathbf{c}^{j_2}$ transforms according to the direct product of the rotational matrices

$$\mathbf{c}^{j_1} \otimes \mathbf{c}^{j_2} \rightarrow \left\{\mathbf{R}^{j_1} \otimes \mathbf{R}^{j_2}\right\} \mathbf{c}^{j_1} \otimes \mathbf{c}^{j_2} = \left\{ \left(\mathbf{H}^{j_1,j_2}\right)^\dagger \left[\bigoplus_{j=|j_1-j_2|}^{j_1+j_2} \mathbf{R}^j \right] \mathbf{H}^{j_1,j_2} \right\} \mathbf{c}^{j_1} \otimes \mathbf{c}^{j_2}.$$

We define $\mathbf{g}^{j_1,j_2,j}$—using the fact that $\mathbf{H}^{j_1,j_2}$ is unitary—as follows:

$$\left[\bigoplus_{j=|j_1-j_2|}^{j_1+j_2} \right] \mathbf{g}^{j_1,j_2,j} \equiv \mathbf{H}^{j_1,j_2} \mathbf{c}^{j_1} \otimes \mathbf{c}^{j_2},$$

which transforms under rotation as $\mathbf{g}^{j,j_1,j_2} \rightarrow \mathbf{R}^j \mathbf{g}^{j_1,j_2,j}$. The cubic rotational invariants, also known as the bispectrum, can be constructed as

$$B_{j_1,j_2,j} = \left(\mathbf{c}^j\right)^\dagger \mathbf{g}^{j_1,j_2,j}.$$

Finally, we arrive to the expression for the bispectrum elements, computed as

$$B_{j_1,j_2,j} = \sum_{m'_1,m_1=-j_1}^{j_1} \sum_{m'_2,m_2=-j_2}^{j_2} \sum_{m',m=-j}^{j} \left(c^j_{m'm}\right)^* C^{jm}_{j_1 m_1 j_2 m_2} C^{jm'}_{j_1 m'_1 j_2 m'_2} c^{j_1}_{m'_1 m_1} c^{j_2}_{m'_2 m_2}.$$

The truncated version of the bispectrum results in a finite array, with 4, 23 and 69 elements for $J_{\max} = 1$, 3 and 5, respectively.

2 Gaussian Process Regression

Notation and formulae:

$$
\begin{aligned}
N &: \text{Number of raw atomic neighbourhood configurations} \\
M &: \text{Number of sparse atomic neighbourhood configurations} \\
\mathbf{x}_n &: \text{bispectrum of } n\text{th reference configuration} \\
\overline{\mathbf{x}}_m &: \text{bispectrum of } m\text{th sparse configuration} \\
\mathbf{x}_* &: \text{bispectrum of configuration for which prediction is sought (“test configuration”)} \\
\mathbf{y} &: \text{vector of data values at the raw configurations} \\
C(\mathbf{x}, \mathbf{x}') &: \text{Covariance function, the measure of similarity of two configurations} \\
[\mathbf{C}_N]_{nn'} &= C(\mathbf{x}_n, \mathbf{x}_{n'}),\ \text{covariance matrix of raw configurations} \\
[\mathbf{C}_M]_{mm'} &= C(\overline{\mathbf{x}}_m, \overline{\mathbf{x}}_{m'}),\ \text{covariance matrix of sparse configurations} \\
[\mathbf{C}_{NM}]_{nm} &= [\mathbf{k}_n]_m = C(\mathbf{x}_n, \overline{\mathbf{x}}_m),\ \text{covariance matrix of sparse and raw configurations, } \mathbf{C}_{MN} = (\mathbf{C}_{NM})^{\mathrm{T}} \\
[\mathbf{k}_*]_m &= C(\overline{\mathbf{x}}_m, \mathbf{x}_*),\ \text{covariance vector of test and sparse configurations} \\
\mathbf{\Lambda} &= \mathrm{Diag}(\mathrm{diag}(\mathbf{C}_N - \mathbf{C}_{NM}\mathbf{C}_M^{-1}\mathbf{C}_{MN})) \\
\sigma &: \text{intrinsic noise of data values (hyperparameter)} \\
\mathbf{Q}_M &= \mathbf{C}_M + \mathbf{C}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{C}_{NM},\ \text{pseudo-covariance matrix of the sparse configurations} \\
\varepsilon_* &= \mathbf{k}_*^T\mathbf{Q}_M^{-1}\mathbf{C}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{y},\ \text{prediction of atomic energy for test configuration} \\
\sigma_*^2 &= C(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{C}_M^{-1} - \mathbf{Q}_M^{-1})\mathbf{k}_* + \sigma^2,\ \text{variance of prediction for test configuration}
\end{aligned}
$$

where $\mathrm{diag}(\mathbf{A})$ is the vector of diagonal elements of the matrix $\mathbf{A}$, and $\mathrm{Diag}(\mathbf{v})$ is the matrix whose diagonal elements are the components of vector $\mathbf{v}$ and the off-diagonal elements are zero.

2.1 Covariance function (kernel)

We use a Gaussian kernel. This kernel enables us to assign a separate *spatial scale* hyperparameter (θ_i) to each element of the feature vector, therefore the kernel provides a more flexible description than a Gaussian kernel with a single hyperparameter. The next three equations show how the covariance function is evaluated if the function values are available for both configurations, if we have the derivative of the function available at one configuration, and if the derivatives are given for both configurations.

$$
\begin{aligned}
C(\mathbf{x}_n, \mathbf{x}_m) &= \delta^2 \exp\left(-\frac{1}{2}\sum_i \frac{(x_n^i - x_m^i)^2}{\theta_i^2}\right) \\
C'(\mathbf{x}_n, \mathbf{x}_m) &= \delta^2 \exp\left(-\frac{1}{2}\sum_i \frac{(x_n^i - x_m^i)^2}{\theta_i^2}\right) \sum_i \frac{x_m^i - x_n^i}{\theta_i^2}\frac{\partial x_m^i}{\partial r_\alpha} \\
C''(\mathbf{x}_n, \mathbf{x}_m) &= \delta^2 \exp\left(-\frac{1}{2}\sum_i \frac{(x_n^i - x_m^i)^2}{\theta_i^2}\right) \left[\sum_i \frac{1}{\theta_i^2}\frac{\partial x_n^i}{\partial r_\alpha}\frac{\partial x_n^i}{\partial r_\beta} - \left(\sum_i \frac{x_n^i - x_m^i}{\theta_i^2}\frac{\partial x_n^i}{\partial r_\alpha}\right)\left(\sum_i \frac{x_n^i - x_m^i}{\theta_i^2}\frac{\partial x_m^i}{\partial r_\beta}\right)\right]
\end{aligned}
$$

2.2 Linear combinations

In our case, only the linear combination of atomic energies can be directly observed. We cannot determine the atomic contributions to the total energy of a system uniquely from an electronic structure calculation. Similarly, the atomic forces—although they are available from first-principles calculations—are not derivatives of atomic energies, but are sums of derivatives of different atomic contributions. It is possible to use a Gaussian Process to infer the underlying function even if only linear combinations of function values are available. Now let $\mathbf{y}$ is the vector of K observed values (total energies and atomic force components). Let $\mathbf{y}'$ be the vector of N *unobserved* values of atomic energies and its derivatives corresponding

to the N atomic neighborhood configurations. Let the $N \times K$ matrix $\mathbf{L}$ describe the relationship of the K observations to the N unknown values. The elements of $\mathbf{L}$ are 0s and 1s, and

$$\mathbf{y} = \mathbf{L}\mathbf{y}'$$

The covariance of the K observations is then given by

$$\mathbf{C}_{KK} = \mathbf{L}^{\mathrm{T}}\mathbf{C}_{NN}\mathbf{L}$$

.

2.3 Putting it all together

The sparsification and the linear combinations are used together to give the final expression for the atomic energy, in such a way that the unobserved values $\mathbf{y}'$ are not needed,

$$\begin{aligned} \mathbf{\Lambda} &= \mathrm{Diag}\left(\mathrm{diag}(\mathbf{L}^T\mathbf{C}_{NN}\mathbf{L} - \mathbf{L}^T\mathbf{C}_{NM}\mathbf{C}_M^{-1}\mathbf{C}_{MN}\mathbf{L})\right) \quad \text{(now a } K \times K \text{ diagonal matrix)} \\ \varepsilon_* &= \mathbf{k}_*^T\left[\mathbf{C}_M + \mathbf{C}_{MN}\mathbf{L}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{L}^T\mathbf{C}_{NM}\right]^{-1}\mathbf{C}_{MN}\mathbf{L}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \end{aligned}$$

3 The data: density functional theory

The DFT data was generated using the local density approximation in case of carbon and the PBE generalized gradient approximation for silicon, germanium, GaN and iron. The electronic Brillouin zone was sampled by using a Monkhorst-Pack k-point grid, with a k-point spacing of at most 0.3Å^{-1} for insulators and 0.14Å^{-1} for iron. The plane wave cutoff was 350, 300, 300, 350, 500 eV for C, Si, Ge, GaN and Fe respectively and the energies were extrapolated to correct for the finite basis set. Ultrasoft pseudopotentials were used with 4 valence electrons for all group IV ions, 3 electrons for Ga, 5 electrons for N and 8 for Fe ions.

4 Testing the GAP parameters

Figure 1 shows the improvement in the distribution of force errors of the GAP model for diamond as the cutoff radius is increased. Figure 2 shows the same as $J_{\max}$ is increased, which corresponds to increasing the spatial resolution of the bispectrum.

5 Potential for gallium nitride

Figures 3 and 4 show the force errors and the phonon spectrum of a simple GAP model for gallium nitride. The long range Coulomb interactions of the ions is significant in this system, so we augmented the local energy of the original GAP model by an Ewald sum of fixed charges (+1 for Ga and -1 for N). The Gaussian Process regression was carried out on forces and energies which were obtained from the DFT calculations by subtracting this Coulomb contribution. The LO/TO splitting in the phonon spectrum shows that the model captures the long range character of the ionic interactions correctly.

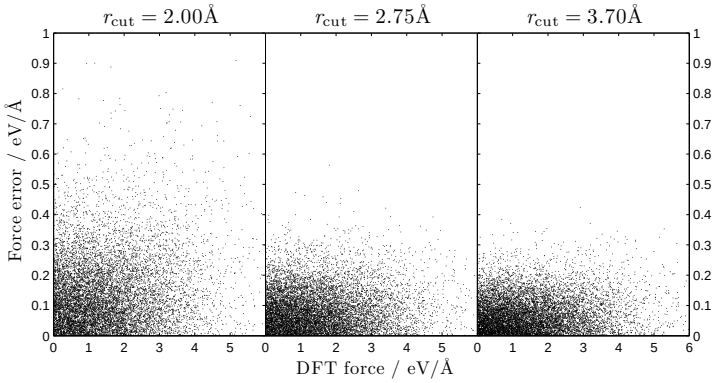

Figure 1: Force correlation of GAP models for diamond with different spatial cutoffs.

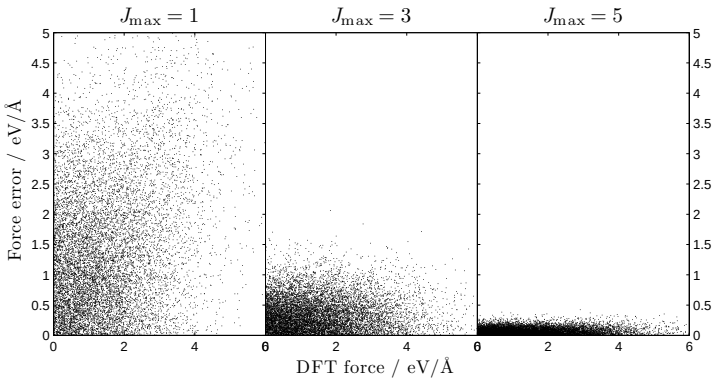

Figure 2: Force correlation of GAP models for diamond with different resolution of representation. The number of invariants were 4, 23 and 69 for J_{max} =1, 3 and 5, respectively.

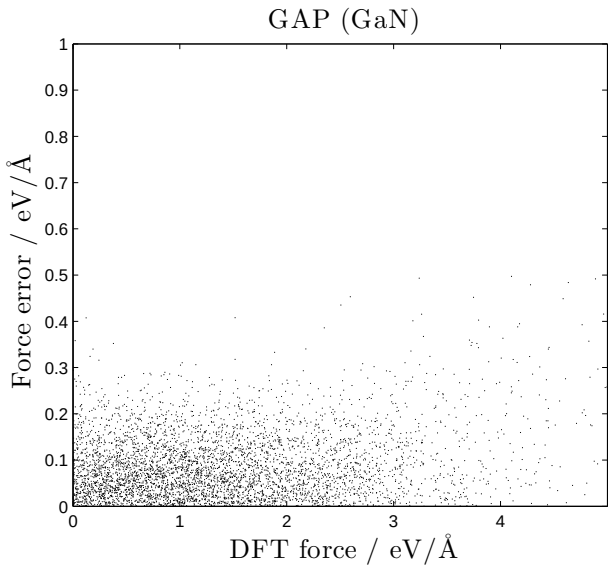

Figure 3: Force errors in GaN of the GAP model augmented with a simple Ewald sum of fixed charges with reference to DFT forces.

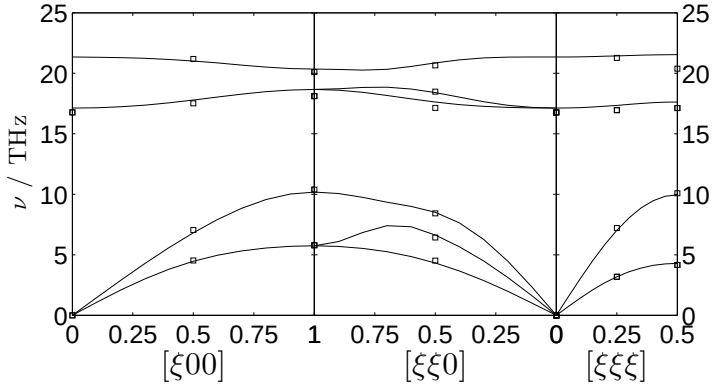

Figure 4: Phonon spectrum of GaN calculated with GAP (solid lines) and PBE-DFT (open squares).